Which strategy to improve the performances of an LSTM-based model for extreme stream temperature values?

by

Mohamed Saadi, Louis Guichard, Gabrielle Cognot, Laurent Labbouz, Hélène Roux

Submitted to *Hydrology and Earth System Sciences*

Manuscript ID: egusphere-2025-3393

Response to Anonymous Referee #2

08 November 2025

1 Summary of main changes

We first would like to thank the Anonymous Referee #2 for their constructive comments, which helped us further verify and clarify our methodological choices. In response to the Referee #2's comments, we made the following main changes:

- We clarified our choice of the potential evapotranspiration as a predictor instead of catchment-scale air temperature (see our response to Comment 1);
- We made additional, computationally expensive tests of new hyperparameter values, which did not substantially improve the test performances (see our response to Comment 4);
- We explained in detail the role of parameters μ and λ of the loss function (see our response to Comments 8, 9 and 10) and underlined, in the manuscript, their importance in emphasizing the weight of extreme stream temperature values in the training phase;
- Finally, we re-organized the subsections of Section 3.2 in response to Comment 7, and updated Figure 5 for clarity.

In the following section, we provide a detailed, point-by-point response to each of the Referee #2's comments.

2 Detailed response to comments from Referee #2

2.1 General and specific comments

General comments: *"This paper compares the performance of LSTM models to predict daily stream water temperature over the whole year, but notably also during the days with the highest 10% of observed water temperatures. Different sets of models are tested based on: (i) local vs regional models, (ii) different sets of input variables, and (iii) different loss functions. Data from several stations in the Garonne River in France are used. The main finding is that regional multi-station training including static attributes improves performance, whereas customized loss functions do not improve performance.*

Overall, the manuscript is well-written, and the figures are clear. The manuscript could be interesting to the readership of HESS. While the knowledge that regional multi-station training including static attributes improves LSTM performance is not very novel, its comparison against the change in performance for different loss functions is valuable. Nevertheless, some important considerations are still needed. Please find below some specific comments and suggestions."

Authors' response: We thank the Referee for their constructive feedback!

Regarding the novelty of our results, we agree that improving the LSTM performances with static attributes and regional training is not novel (e.g., Kratzert et al., 2024). However, our setup quantifies the contribution of including static attributes in comparison to just training the LSTM at many stations, which helps highlight the mechanisms by which LSTM models perform better when trained regionally (see for example, Yu et al., 2024). In fact, a simple regional training of LSTM models is detrimental in terms of test performances (for instance, compare the results shown in Figure 3 with those shown in Figure 4 with only dynamic attributes). Only when static attributes are included that regional training becomes more efficient than local training.

Regarding the specific comments and suggestions, below we provide a detailed answer.

Comment 1: *"The computation of catchment average potential evapotranspiration (PE) according to Oudin et al. (2005) seems unnecessary. Catchment average Ta and information on day of the year could be used instead of PE, unless PE is strictly necessary for obtaining Qsim."*

Authors' response: Catchment-scale potential evapotranspiration (PE) is one of the required inputs to the hydrological model GR6J that we used to reconstruct streamflow (Q_{sim}) at the location of stream temperature stations. In the PE formula by Oudin et al. (2006), PE is (almost) linearly dependent on catchment-average air temperature, meaning that the use of PE as input to the LSTM model provides almost the same information content as that of catchment-average air temperature. Furthermore, we wanted to compare (1) the option of letting the LSTM model decide which information to extract from precipitation (P) and PE that is most relevant to the reproduction of stream temperature, against (2) the option of restraining this hydrological input to the simulated streamflow Q_{sim} only, which somewhat represents a hydrologically digested version of P and PE, and also a more relevant hydrological variable (since it is at the station scale) than P and PE (which are at the catchment scale). Figures 3 and 4 show that these two options provide comparable or similar performances, highlighting their equivalence for the reproduction of daily stream temperature values.

As for the use of the day of the year, we intentionally excluded any features explicitly based on time (day or month of the year), which were used e.g. by Feigl et al. (2021). We fear that the inclusion of these time-based features might overshadow the importance of more physically relevant variables (namely station-scale air temperature and hydrological variables),

knowing that stream temperature is a variable that features a strong seasonal variability. In addition, these time-based features are generally not used as forcing by process-based models, which would hinder a fair comparison between LSTM-based models and process-based models, or could suggest that the better performances of LSTM-based models are attributable to the "excessive" reliance on feature engineering. We clarified these choices in the revised manuscript version, as we stated by the end of Section 3.2.3 of the revised manuscript that:

"Finally, we avoided using time-based features (month or day of the year; Feigl et al., 2021) so that the performances of the tested LSTM models remain comparable to process-based models that do not benefit from feature engineering. In addition, knowing the strong seasonality of T_w and of some of the input variables (T_a, T_{amn}, T_{amx}, and PE), the use of time-based features would be redundant information-wise and would likely lead to gains in predictive performances high enough to overshadow the contributions of the more physically relevant variables used in our setup."

Comment 2: *"It could be useful to have an additional table showing the values from the different input variables for all stations, even if it is in the appendix."*

Authors' response: We believe that the Referee's comment refers to static attributes, as including the values of dynamic variables is unpractical. For static attributes, we believe that Table 1 provides a concise and fairly informative description of the richness of our dataset using the range (min and max values) and the median values of the geographical and climatic features of our catchment set. Adding a fourth appendix to show the values for each one of the 37 stations would only unnecessarily increase the length of the manuscript. Note that these data are provided in the Zenodo repository (https://zenodo.org/records/15864784, file "`data/2024-09-09_ListStationsTw_StaticDesc_v03.TXT`") accompanying the manuscript submission that contains all necessary scripts used to run our experiments (Saadi, 2025).

Comment 3: *"It could be useful to include the long-term mean and standard deviation of daily water temperature from each station as a static variable, but I understand this might not be feasible if it means that all models need to be re-trained."*

Authors' response: To test the contribution of an additional static attribute we would have to re-run all the 378 ($7 \times 18 \times 3$) regional models, which is computationally very expensive (see Appendix A of the original manuscript for wall-clock times needed for training). Adding static attributes that are computed from the target variable that we want to reproduce by the LSTM is, in our sense, would result in an implicit information leakage, thus resulting in (a priori) better performing but less robust LSTM models. In addition, this will substantially hinder any application of the regionally trained models at ungauged locations, where stream temperature records (hence any statistic of stream temperature needed for that kind of model applications) is absent.

Comment 4: *"L211: It would be useful to do a more detail assessment for choosing the hyperparameters of the LSTM models, considering the findings of Feigl et al. (2021). Doing hyperparameter optimization as in Kraft et al. (2025) would be a good option.*

Kraft, B., Schirmer, M., Aeberhard, W. H., Zappa, M., Seneviratne, S. I., and Gudmundsson, L.: CH-RUN: a deep-learning-based spatially contiguous runoff reconstruction for Switzerland, Hydrol. Earth Syst. Sci., 29, 1061–1082, https://doi.org/10.5194/hess-29-1061-2025, 2025."

Authors' response: We agree with the Referee that it would be useful to do a more detailed assessment of the effect of hyperparameters (number of layers, number of cells per layer,

batch size, etc.), but we already found in our preliminary tests that these have little effect on model performances. Thus, we made choices that are in line with previous studies (Hashemi et al., 2022; Kratzert et al., 2019; Rahmani et al., 2021a, 2021b). These choices are also in line with the best choices found by Kraft et al. (2025).

In our response to Comment 9 made by Referee #1, we made additional, computationally expensive tests using a higher learning rate of 10^{-3} and a dropout rate at 0.1 with the following configurations:

- Two values for the number of layers (NL): 1 and 2;
- Two values for the number of cells per layer (HDN_SZ): 128 and 256;
- Two values for the batch size (BTH_SZ): 64 and 256;
- Four loss functions corresponding to μ = 1, λ = 1, 2, with and without standardization of the target variable;
- Six input-variable sets: T_{amn}+T_{amx}, T_{amn}+T_{amx}+CatAttrs, T_{amn}+T_{amx}+P+PE, T_{amn}+ T_{amx}+P+PE+CatAttrs, T_{amn}+T_{amx}+Q_{sim}, and T_{amn}+ T_{amx}+Q_{sim}+CatAttrs.
- Three values for lookback: 30, 90, and 365.
- All models are trained regionally.

This amounted to training 2 × 2 × 2 × 4 × 6 × 3 = 576 models using the same methodological choices as in our main study. Note that we kept only the best lookback value based on the MSE of the validation period, hence showing the test performances for only 192 models. Figure R1 shows that tuning any of the three hyperparameters (number of layers, number of cells per layer or the batch size) has negligible effect on model performances over the whole test period (Figures R1a to R1c) and over the top 10% values of the test period (Figures R1d to R1f).

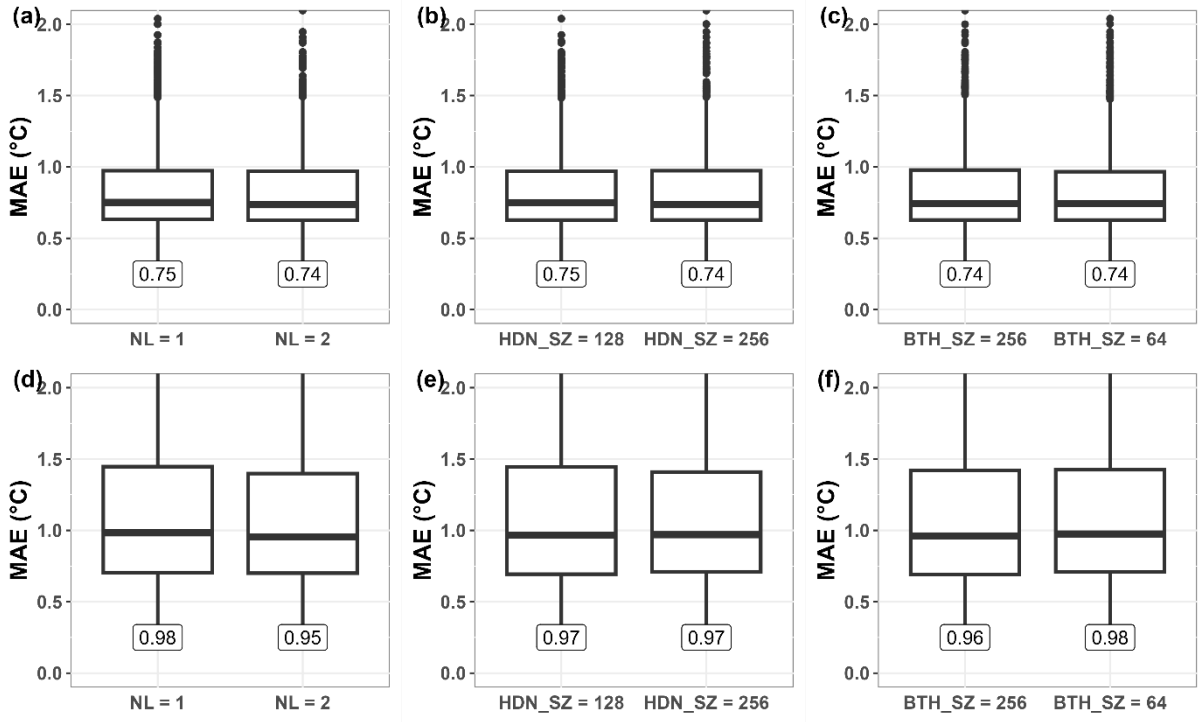

Figure R1: Effect of the number of layers (NL), the number of cells per layer (HDN_SZ), and the batch size (BTH_SZ) on model performances over (a)-(c) the whole test period, and (d)-(f) the top 10% values of the test period. Values under each box indicate the median value. Each distribution is computed from a set of 2016 points. Note that these model runs are made with a learning rate at 10^{-3} and a dropout rate at 0.1.

Finally, implementing the hyperparameter tuning experiment by Kraft et al. (2025) is computationally challenging in our case, noting that we tested 18 loss functions (compared to one loss function in Kraft et al. (2025)), several choices of the input variables, and local vs. regional training.

In the revised manuscript version, we cited the study of Kraft et al. (2025) as part of the very recent literature that applied LSTM models, and also to highlight their method in embedding static features for regional training, both in Section 1.2:

"Among these techniques, models based on LSTM (Long Short-Term Memory; Hochreiter and Schmidhuber, 1997) have demonstrated excellent performances in predicting not only stream temperature but also several other dynamic, environmental variables (Arsenault et al., 2023; Kraft et al., 2025; Kratzert et al., 2018; Ma et al., 2021; Nearing et al., 2024; Song et al., 2024; Zhi et al., 2021)."

and in Section 3.2.3 of the revised manuscript to highlight their assessment of different strategies of embedding static attributes as inputs to LSTM models:

"Note that to feed the LSTM model with the static attributes, we opted for a simple integration strategy (see, e.g., Hashemi et al., 2022) in which we repeated the value of each static attribute at each time step to match the length of the dynamic attributes, then we concatenated the columns of the static attributes to those of the dynamic attributes (for each catchment). This strategy compared well against a separate processing of static attributes from dynamic ones using an entity-aware (EA) variant of LSTM networks (Kratzert et al., 2019), and better strategies to encode the static attributes as well as the dynamic variables as inputs to LSTM models have been recently intercompared by Kraft et al. (2025)."

Comment 5: *"An important point when training deep learning models is their inherent randomness. It would be useful to assess for each model setup the variability in the performance when retraining the model with different random seeds. In this way, the differences in performance from the different strategies tested in the paper can be put into context with the uncertainty in performance from varying random seeds."*

Authors' response: We agree with the Referee's comment regarding the inherent randomness in terms of model performances due to the randomly assigned initial values of model parameters prior to training. However, we have several reasons not to make a detailed assessment of the effect of this randomness on our conclusions, as requested by the Referee.

First, there is no reason for this randomness to introduce a bias in favour of one of the options that we tested. For example, in Figure 2, for each loss function, median statistics were computed from $4 \times 21 = 84$ locally trained models (Figure 2a) and $7 \times 1 = 7$ regionally trained models (Figure 2b), meaning that the variability in LSTM performances due to random initialization is already assessed for each loss function thanks to these repetitions induced by several choices of input variables, under the (most likely valid) assumption that the effect of model initialization is independent from the choice of the input variables. The same can be said for the comparison of the different choices of input variables for locally trained models and regionally trained models shown in Figures 3 and 4 of the manuscript.

Second, we used a binomial test to assess the significance of the differences in performances between the different sets of options (Fidal and Kjeldsen, 2020; Saadi et al., 2021). This binomial test does not specifically look at the magnitude of improvements, but at the number of times an option A (say a loss function or a set of input variables) performs better than another option B. When option B performs systematically better than option A (hence a significant binomial test), there is strong reason to believe that option B brings significant improvements compared to option A. We applied this statistical test to compare all our options, and letters in

Figures 3 and 4 for example show that the use of hydrological variables and static attributes significantly improves the LSTM performances (in the sense of the binomial test).

Third, it would take us months to provide, for each model setup, a detailed assessment of the variability in terms of model performances due to random initial parameter values (cf. the wall-clock times needed to train local and regional models in Appendix A of the original manuscript version). In a preliminary work, we attempted at getting an order of magnitude of the changes in model performances due to this random initialization of model parameters. We retrained 100 times a regionally trained model on 26 stations with a sequence length of 16 days. Note that these choices are way different than the final setup used in our study. Test results for 10 stations show that the standard-deviation can be up to 0.07°C, with fluctuations (i.e., difference between min and max performances) reaching up to 0.35°C in MAE, as can be seen in Figure R2. The improvements that the static attributes brought in in terms of median MAE are way larger than these values, as can be seen in Figure 4 of the original manuscript.

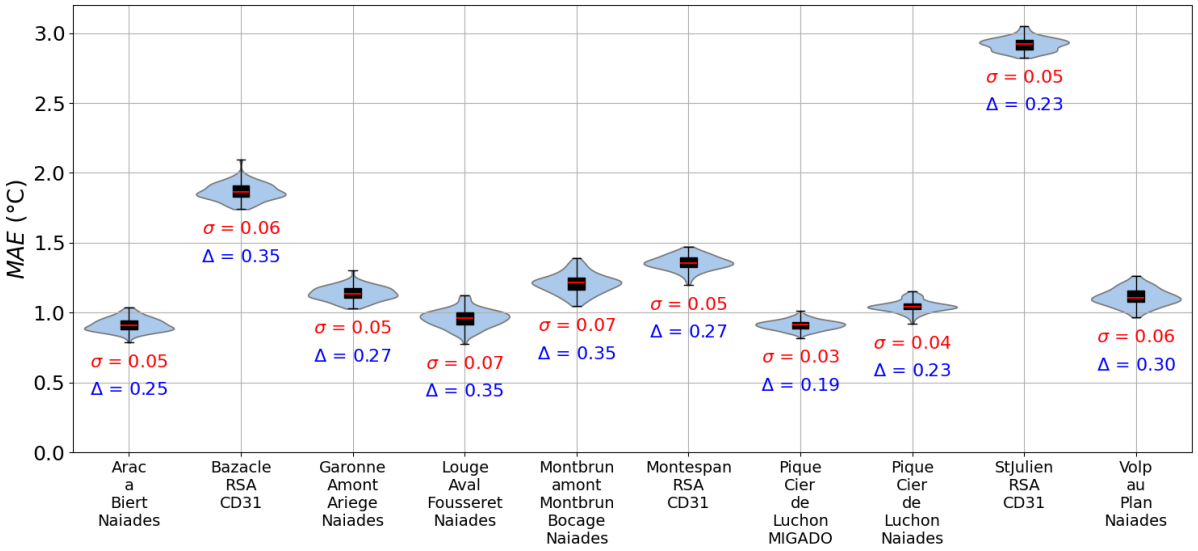

Figure R2: Effect of random parameter initialization on model performances using an ensemble of 100 regionally trained models. For each distribution, values in red refer to the standard deviation, and values in blue refer to the range (max-min).

Comment 6: *"L240–244: I would suggest constructing the validation set using 15% of the observations at all stations. Using more data from the non-test stations could bias the models to better fit to these stations instead of to the test stations. Please also clarify if this 15% corresponds to at least one continuous year of observations."*

Authors' response: We believe that our choice of using 30% of the observations from the non-test stations instead of 15% adds only to the robustness of our regionally trained models. In addition, the total number of datapoints from the 21 test stations is 91705 days, of which 13755 (15% of 91705) are used for validation. The total number of datapoints from the non-test stations is 12626, of which 3788 (30% of 12626) are used for validation. This means that in our configuration, datapoints from the test stations constitute up to 78% (13755/(13755+3788)) of the total number of points on which the regional models are validated. Lowering the contribution of the non-test stations from 30% to 15% would only lead to increasing the dominance of test stations in the validation data from 78% to 88%, and unnecessarily decreasing the total number of points on which the regional models are validated. Finally, the length of available records of stream temperature in non-test stations ranges from 73 to 2365 days, and these datapoints do not span a continuous year for at least half of the non-test stations.

We modified Lines 240-244 of the original manuscript version to emphasize that the test stations still dominate the validation set. Now this part of the text reads:

"[…] The set of regional models trained using data from all the 37 stations. In this case, we constructed the training set by concatenating 70% of the available T_w and their corresponding input variables from each station. The validation set was constructed using 15% of observations at the test stations (21/37) and the whole remaining 30% of observations at the non-test stations (16/37). Finally, the remaining 15% of observations at the test stations were used to test the regional models, thus enabling a comparison of locally and regionally trained models on the same datapoints. Although 30% of the observations at the non-test stations are used in the validation set against 15% from the test stations, datapoints from the test stations still constitute up to 78% of the validation set due to the low availability of T_w records at non-test stations."

Comment 7: *"Suggestion to move section 3.2.3 to 3.2.1 to be more consistent with the order proposed in the last paragraph of section 1 and in section 4."*

Authors' response: Although this is not necessary to follow the methodological setup, we moved Section 3.2.3 of the original manuscript to 3.2.1, Section 3.2.1 to 3.2.2, and Section 3.2.2 to 3.2.3. The introductory paragraph of Section 3.2 now summarizes this part as follows:

"In this part, we summarize the strategies that we tested to look for the best approach to improve the LSTM performances at extreme, high T_w values. We define these values as the daily (average) T_w values exceeded less than 10% of the time. Our tested strategies include an adaptation of the loss function to increase the weight of extreme values in the training phase (Sect. 3.2.1), regional multi-catchment training (Sect. 3.2.2), and the inclusion of hydrologically relevant variables and static attributes (Sect. 3.2.3)."

Comment 8: *"Important: I don't see the need to have the denominator in Eq. 3, and it seems to be counterproductive. When having high Tw and u, the denominator increases faster the numerator, thus reducing the loss for higher values of Tw (see Table below with example data). If this is the case, then the loss function does not serve its intended purpose to give higher weights to errors when Tw values are high. I think only the numerator of Eq. 3 should be used as loss function."*

u =	3				
lambda =	2				
T_w_obs	T_w_sim	T_w_obs_bar	Eq. 3 numerator	Eq. 3 denominator	Loss
20	19.5	13	342371.2656	33674809	0.010166985
15	14.5	13	106520.6406	1387684	0.076761453
10	9.5	13	20341.89063	1432809	0.01419721

Authors' response: Looking at the "Loss" column, we believe that the Referee is not computing exactly the same loss function as ours: Instead of computing the ratio of the sums of errors $\frac{\sum_t \left|T^{\mu}_{w,obs,t} - T^{\mu}_{w,sim,t}\right|^{\lambda}}{\sum_t \left|T^{\mu}_{w,obs,t} - \overline{T^{\mu}_{w,obs}}\right|^{\lambda}}$, the Referee computed point-wise fractions of errors at each point, i.e., $\frac{\left|T^{\mu}_{w,obs,t} - T^{\mu}_{w,sim,t}\right|^{\lambda}}{\left|T^{\mu}_{w,obs,t} - \overline{T^{\mu}_{w,obs}}\right|^{\lambda}}$, and of course these ratios are higher for points closer to the mean of observations than points farther from the mean, like the extreme values. An alternative, more rigorous way of evaluating the relative importance of datapoints in the training phase is to

compute the magnitude (absolute value) of the sensitivity of the loss function to the simulated stream temperature value, which gives in our case:

$$\left|\frac{\partial \mathcal{L}}{\partial T_{w,sim,t}}\right| = \frac{\lambda\mu T_{w,sim,t}^{\mu-1}\left|T_{w,sim,t}^{\mu} - T_{w,obs,t}^{\mu}\right|^{\lambda-1}}{\sum_t \left|T_{w,obs,t}^{\mu} - \overline{T_{w,obs}^{\mu}}\right|^{\lambda}}$$

Let us consider the time steps H (for high) and L (for low) where a high, extreme stream temperature $T_{w,obs,H}$ and a low stream temperature $T_{w,obs,L}$ are observed. The ratio of the absolute sensitivity values of the loss function to model simulations at high vs. low stream temperature observation time steps can be written as

$$r_{H/L} = \left(\frac{T_{w,sim,H}}{T_{w,sim,L}}\right)^{\mu-1}\left|\frac{T_{w,sim,H}^{\mu} - T_{w,obs,H}^{\mu}}{T_{w,sim,L}^{\mu} - T_{w,obs,L}^{\mu}}\right|^{\lambda-1}$$

We can see that the denominator plays only the role of a normalizing constant in the loss function, and does not affect the relative importance of model simulations during the training phase because it cancels out in $r_{H/L}$. If we consider the general case, $r_{H/L}$ is proportional to the errors of the simulations when $\lambda > 1$ (i.e., for MSE and M4E in our case). But to understand the weight of extreme vs. usual values (i.e., values closer to the mean), we can consider a simple case where we have a constant shift Δ between the observations and the simulations, i.e., that $T_{w,sim,t} = T_{w,obs,t} + \Delta$ for all time steps t. In this case, normally, if the loss function is somewhat "egalitarian", the ratio should be close to 1; if it emphasizes extreme large values, it should be much higher than 1, and if it emphasizes low values, it should be much lower than 1. The expression of the ratio becomes

$$r_{H/L} = \left(\frac{T_{w,obs,H} + \Delta}{T_{w,obs,L} + \Delta}\right)^{\mu-1}\left|\frac{\left(T_{w,obs,H} + \Delta\right)^{\mu} - T_{w,obs,H}^{\mu}}{\left(T_{w,obs,L} + \Delta\right)^{\mu} - T_{w,obs,L}^{\mu}}\right|^{\lambda-1}$$

A numerical application for $\Delta = 1°C$, $T_{w,obs,L} = 15°C$ and $T_{w,obs,H} = 30°C$ gives the ratios in Table R1. We can see that with $\mu = 1$, there is in this simple case no overweighting of the extreme value $T_{w,obs,H}$ compared to the low value $T_{w,obs,L}$. As λ and especially as μ increases, the relative importance of the extreme value compared to the average value skyrockets.

Table R1: Values for the ratio $r_{H/L}$ with $\Delta = 1°C$, $T_{w,obs,L} = 15°C$ and $T_{w,obs,H} = 30°C$ computed for all combinations of λ and μ values tested in our study.

	μ = 1	μ = 3	μ = 5
λ = 1 (MAE)	1	4	14
λ = 2 (MSE)	1	15	211
λ = 4 (M4E)	1	218	47269

In summary, the denominator does not theoretically impact the relative importance of the extreme values in the loss function and plays only the role of a normalizing constant. In our opinion, it's important to keep this denominator because (1) it helps interpret the loss function, and (2) without this denominator the learning rate should be modified to account for the large magnitudes of errors with larger values of μ (and λ). We added a sentence in the revised manuscript to justify this choice:

"For an intuitive interpretation of this function, the denominator in the loss function of Eq. (1) standardizes the values of the loss function by comparing the performances of the LSTM model to a "dummy" model that predicts the average value of the transformed observations $\overline{g(T_{w,obs})^{\mu}}$ for all the time steps. This denominator plays also the role of a

normalizing constant without which the learning rate should be adapted to account for the larger error magnitudes in the numerator induced by higher values of λ and especially μ. This last hyperparameter μ results in magnitudes of $g(T_w)$ that are higher at extreme values than at mild values, inducing larger errors at (and thus more emphasis on) extremely high values."

Comment 9: *"L282: Explain what u does in Eq. 3, i.e. having higher powers on higher Tw values would lead to larger errors, thus emphasizing the weight on high Tw, if I understood it correctly."*

Authors' response: Exactly! We give more explanations regarding the role of μ (and λ) in our answer to Comment 8.

In the manuscript, we added a sentence to give an interpretation of the role of μ in Equation 3 as suggested by the Referee:

"For an intuitive interpretation of this function, the denominator in the loss function of Eq. (1) standardizes the values of the loss function by comparing the performances of the LSTM model to a "dummy" model that predicts the average value of the transformed observations $\overline{g(T_{w,obs})}^{\mu}$ for all the time steps. This denominator plays also the role of a normalizing constant without which the learning rate should be adapted to account for the larger error magnitudes in the numerator induced by higher values of λ and especially μ. This last hyperparameter μ results in magnitudes of $g(T_w)$ that are higher at extreme values than at mild values, inducing larger errors at (and thus more emphasis on) extremely high values."

Comment 10: *"L287–289: This is important. It would be useful to add another sentence or example to clarify that having higher powers on higher Tw values would lead to larger errors, thus emphasizing the weight on high Tw."*

Authors' response: In response to Comment 9, we added a sentence that helps understand the role of μ in the loss function. Please see our answer to Comment 9 and also Comment 8 for a numerical illustration of the role of μ (and also λ) in the loss function.

Comment 11: *"L324: Report the number of cases out of the 21 for which the performance improved. This is more informative than saying it is not statistically significant."*

Authors' response: First, the total number of cases is not 21, but 21 times the number of sets of input variables, which is 4 for the local models, and 7 for the regional models. This is why in Figure 2, we represented the percentage of cases for which the use of a loss function improves on the use of the reference loss function (MSE on standardized target). Anyway, we added the number of cases when commenting Figure 2 to better understand the results of the statistical test. This part now reads:

"[…] In detail:

- ***The best performances over the whole period were systematically obtained using MAE (with or without standardization) as a loss function in the training: median test MAE reached 0.72°C for the local models (all configurations of input variables combined, Fig. 2a) and 0.77°C for the regionally trained models (Fig. 2b). In comparison to the reference loss function (MSE on standardized target), these improvements with MAE as a loss function were statistically significant only in the case of locally trained models, for which MAE applied to standardized and non-standardized target resulted in better performances than the reference loss function for 58/84 and 54/84 cases, respectively (recall that the number of***

cases is the number of the test stations times the number of sets of input variables, see Table 2). For the regional models, MSE without standardization is the only loss function that resulted in statistically better performances, with better scores than the reference loss function for 85/147 cases, which is higher in this case than the 5%-significance threshold (84/147).

- ***Compared to the whole-range performances, model performances were systematically lower on the top 10% range. For this range, the best median performances were obtained using MAE with standardization for the local models (1.07°C, Fig. 2a) or using MSE without standardization for the regional models (0.98°C, Fig. 2b). However, these improvements were not statistically significant at the 5%-level in comparison with the reference loss function (MSE with standardized target): For the local models, MAE with standardization resulted in better performances compared to the reference loss function for only 48/84 cases, which is below the 5%-significance threshold (51/84); For the regional models, MSE without standardization gave better scores than the reference loss function for only 70/147 cases."***

2.2 Minor comments and technical corrections

Comment 1: *"The study from Padrón et al. (2025) could be useful for section 1.2 and the second paragraph of section 5.*

Padrón, R. S., Zappa, M., Bernhard, L., and Bogner, K.: Extended-range forecasting of stream water temperature with deep-learning models, Hydrol. Earth Syst. Sci., 29, 1685–1702, https://doi.org/10.5194/hess-29-1685-2025, 2025."

Authors' response: We thank the Referee for this relevant and timely suggestion. We added a citation of the work of Padrón et al. (2025) by the end of Section 1.2 to cite examples oriented towards forecasting tasks:

"Other studies demonstrated successful applications of LSTM networks for the more operational task of stream temperature forecasting (e.g., Padrón et al., 2025; Qiu et al., 2021; Zwart et al., 2023)."

Comment 2: *"Table 1: Please clarify what are the min, median and max values reported. Are these average values across all 21 stations? Otherwise, it is not consistent with the value of 21C reported in L381."*

Authors' response: In Line 381 of the manuscript, these values represent the min-max values for the station in question (Garonne at Valentine). The values min-max of Table 1 represent the min-max of the long-term average of stream temperature. In other words:

- We computed the long-term average of stream temperature for each station, i.e., the average of all records. This provides a sample of 37 values (for 37 stations);
- Since some stations have a high rate of missing values, the averages may not be representative. For this reason, we excluded the 16 non-test stations, which have a number of available stream temperature observations lower than 2434 days;
- We finally calculated the min, max and median values using the remaining 21 values.

To add these clarifications, we modified the title of Table 1 as follows:

"Table 2: Summary of dynamic and static variables and their distributions across the catchment set. Min, median, and max values were computed from the set of 37 stations, except for the long-term averages of daily stream temperature."

Additionally, at the bottom of Table 1, we added the following note regarding the values for daily stream temperature:

"[a]We first computed, for each station, the long-term average of stream temperature values using the whole time series, which gave 37 values. We then excluded the stations with short time series by computing min, median, and max statistics using the 21 test stations only."

Comment 3: *"L238–239: Mention here that the 15% of the available records span at least one full year."*

Authors' response: We now mention in those lines that 15% of the available records represent at least one-year worth of observations:

"[…] The set of local models trained using the first 70% of the available T_w records and their corresponding input, dynamic variables only at the station of interest. In this case, half of the remaining T_w observations (i.e., 15% of the available records) were used for validation and the remaining records (i.e., 15% of the available records) were kept for test. Note that for these stations, 15% of the available records span at least one-year worth of daily observations."

Comment 4: *"Figs. 3 and 4: Clarify if the bottom row corresponds to the best loss function averaged over all sets of input variables. If this is not the case, then why is the MAE of the "Reference" loss function (1.29) lower than that of the "Best" loss function (1.48) for the model with only Ta as input in Fig. 4."*

Authors' response: Yes, the bottom row of Figures 3 and 4 corresponds to the best loss function defined as the one with the best median value over all sets of input variables. We now clarified this in the captions of Figures 3, 4, C2, and C3 of the revised manuscript. For example, the caption of Figure 3 now reads:

"Figure 3: Distributions of the test performances (MAE, in °C) of the local models over the whole test period (left column) and the period corresponding to the highest 10% observed values (right column). The top row shows the distribution over all the loss functions (18 × 21 = 378 points per distribution). The middle row shows the performances for the reference loss function (MSE with standardization, 21 points per distribution). The bottom row shows the performances for the best loss function over all sets of input variables (MAE without standardization for the whole period and MAE with standardization for the top 10%, 21 points per distribution). Numerical values under the boxes represent the median value for each distribution. Letters under the numerical values rank the distributions, and are defined such as distributions that share at least one letter are not significantly different according to the binomial test."

Comment 5: *"L381: Should "which" be replaced by "with"."*

Authors' response: We thank the Referee for underlining this typo, which we have corrected in the revised version.

Comment 6: *"Fig. 5: suggestion to reduce the size of the black dots to improve visualization."*

Authors' response: We now reduced the size of the black dots to improve the visualization of Figure 5 (see Figure R3).

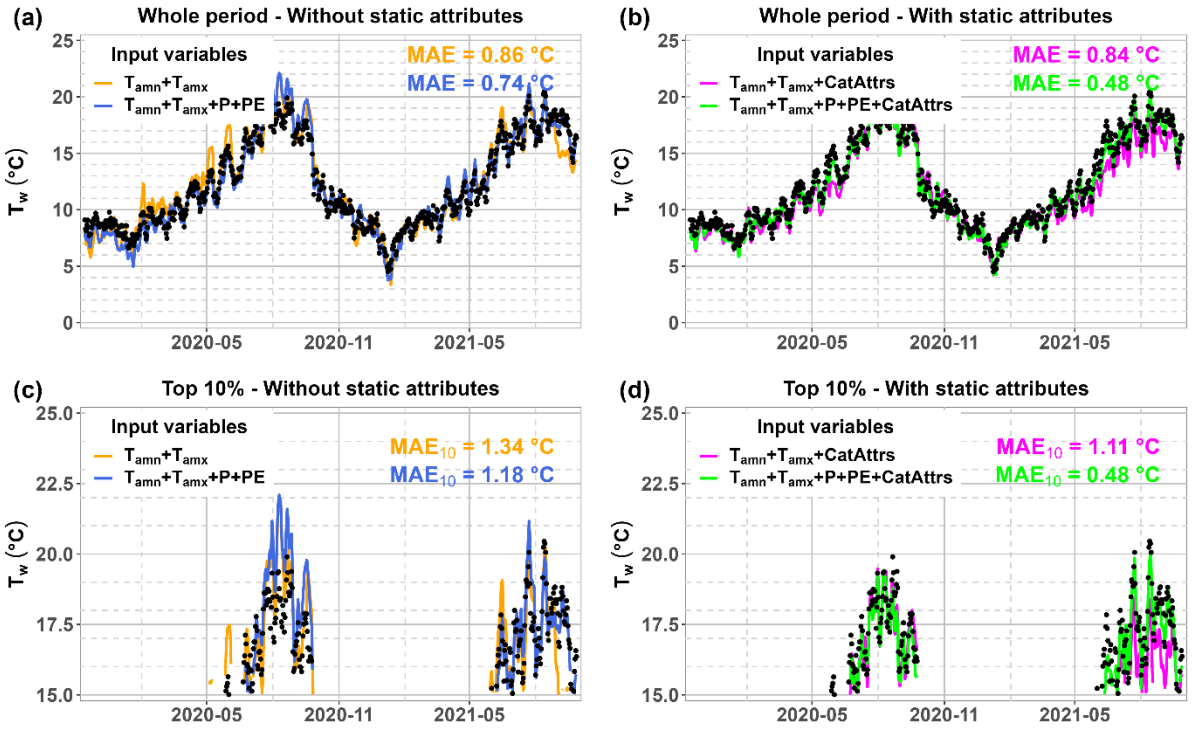

Figure R3: Revised Figure 5 of the manuscript, which shows an example of model simulations using LSTM models against observations at the station of the Garonne at Valentine (MIGADO). All models are regionally trained. The two top figures (a and b) show the model simulations and performances (MAE, in °C) over the whole test period, and the two bottom figures (c and d) zoom in on the highest 10% of the observations during the test period. Input variables include station-scale daily minimum and maximum air temperatures (T_{amn}+T_{amx}) in addition to catchment-scale precipitation and potential evapotranspiration (P+PE). Models that use static attributes among input variables are shown on the right (b and d), while models that use only dynamic variables are shown on the left (a and c).

Comment 7: *"Fig. 5 caption: "(a and c)" should be exchanged with "(b and d)" and vice versa."*

Authors' response: We thank the Referee for mentioning this typo, which we have corrected in the revised manuscript version.

3 Cited References

Arsenault, R., Martel, J.-L., Brunet, F., Brissette, F., Mai, J., 2023. Continuous streamflow prediction in ungauged basins: long short-term memory neural networks clearly outperform traditional hydrological models. Hydrol. Earth Syst. Sci. 27, 139–157. https://doi.org/10.5194/hess-27-139-2023

Feigl, M., Lebiedzinski, K., Herrnegger, M., Schulz, K., 2021. Machine-learning methods for stream water temperature prediction. Hydrol. Earth Syst. Sci. 25, 2951–2977. https://doi.org/10.5194/hess-25-2951-2021

Fidal, J., Kjeldsen, T.R., 2020. Operational comparison of rainfall-runoff models through hypothesis testing. J. Hydrol. Eng. 25, 04020005. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001892

Hashemi, R., Brigode, P., Garambois, P.-A., Javelle, P., 2022. How can we benefit from regime information to make more effective use of long short-term memory (LSTM) runoff models? Hydrol. Earth Syst. Sci. 26, 5793–5816. https://doi.org/10.5194/hess-26-5793-2022

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Comput. 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Kraft, B., Schirmer, M., Aeberhard, W.H., Zappa, M., Seneviratne, S.I., Gudmundsson, L., 2025. CH-RUN: a deep-learning-based spatially contiguous runoff reconstruction for Switzerland. Hydrol. Earth Syst. Sci. 29, 1061–1082. https://doi.org/10.5194/hess-29-1061-2025

Kratzert, F., Gauch, M., Klotz, D., Nearing, G., 2024. HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin. Hydrol. Earth Syst. Sci. 28, 4187–4201. https://doi.org/10.5194/hess-28-4187-2024

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrol. Earth Syst. Sci. 22, 6005–6022. https://doi.org/10.5194/hess-22-6005-2018

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., Nearing, G., 2019. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrol. Earth Syst. Sci. 23, 5089–5110. https://doi.org/10.5194/hess-23-5089-2019

Ma, Y., Montzka, C., Bayat, B., Kollet, S., 2021. Using Long Short-Term Memory networks to connect water table depth anomalies to precipitation anomalies over Europe. Hydrol. Earth Syst. Sci. 25, 3555–3575. https://doi.org/10.5194/hess-25-3555-2021

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C., Shalev, G., Shenzis, S., Tekalign, T.Y., Weitzner, D., Matias, Y., 2024. Global prediction of extreme floods in ungauged watersheds. Nature 627, 559–563. https://doi.org/10.1038/s41586-024-07145-1

Oudin, L., Andréassian, V., Loumagne, C., Michel, C., 2006. How informative is land-cover for the regionalization of the GR4J rainfall-runoff model? Lessons of a downward approach. IAHS Publ. 307, 246–255.

Padrón, R.S., Zappa, M., Bernhard, L., Bogner, K., 2025. Extended-range forecasting of stream water temperature with deep-learning models. Hydrol. Earth Syst. Sci. 29, 1685–1702. https://doi.org/10.5194/hess-29-1685-2025

Qiu, R., Wang, Y., Rhoads, B., Wang, D., Qiu, W., Tao, Y., Wu, J., 2021. River water temperature forecasting using a deep learning method. J. Hydrol. 595, 126016. https://doi.org/10.1016/j.jhydrol.2021.126016

Rahmani, F., Lawson, K., Ouyang, W., Appling, A., Oliver, S., Shen, C., 2021a. Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data. Environ. Res. Lett. 16, 024025. https://doi.org/10.1088/1748-9326/abd501

Rahmani, F., Shen, C., Oliver, S., Lawson, K., Appling, A., 2021b. Deep learning approaches for improving prediction of daily stream temperature in data-scarce, unmonitored, and dammed basins. Hydrol. Process. 35, e14400. https://doi.org/10.1002/hyp.14400

Saadi, M., 2025. Scripts for the paper “Which strategy to improve the performances of an LSTM-based model for extreme stream temperature values?” https://doi.org/10.5281/zenodo.15864784

Saadi, M., Oudin, L., Ribstein, P., 2021. Physically consistent conceptual rainfall–runoff model for urbanized catchments. J. Hydrol. 599, 126394. https://doi.org/10.1016/j.jhydrol.2021.126394

Song, Y., Chaemchuen, P., Rahmani, F., Zhi, W., Li, L., Liu, X., Boyer, E., Bindas, T., Lawson, K., Shen, C., 2024. Deep learning insights into suspended sediment concentrations

across the conterminous United States: Strengths and limitations. J. Hydrol. 639, 131573. https://doi.org/10.1016/j.jhydrol.2024.131573

Yu, Q., Jiang, L., Schneider, R., Zheng, Y., Liu, J., 2024. Deciphering the Mechanism of Better Predictions of Regional LSTM Models in Ungauged Basins. Water Resour. Res. 60, e2023WR035876. https://doi.org/10.1029/2023WR035876

Zhi, W., Feng, D., Tsai, W.-P., Sterle, G., Harpold, A., Shen, C., Li, L., 2021. From Hydrometeorology to River Water Quality: Can a Deep Learning Model Predict Dissolved Oxygen at the Continental Scale? Environ. Sci. Technol. 55, 2357–2368. https://doi.org/10.1021/acs.est.0c06783

Zwart, J.A., Oliver, S.K., Watkins, W.D., Sadler, J.M., Appling, A.P., Corson-Dosch, H.R., Jia, X., Kumar, V., Read, J.S., 2023. Near-term forecasts of stream temperature using deep learning and data assimilation in support of management decisions. JAWRA J. Am. Water Resour. Assoc. 59, 317–337. https://doi.org/10.1111/1752-1688.13093